$$E_{pair} = \Sigma\Sigma E_{ij} \tag{8a}$$

where:

$$E_{ij} = \begin{cases} \infty, & \text{for } r_{ij} < 3 \\ E^{rep}, & \text{for } 3 \le r_{ij} < R_{i,j}^{rep} \\ \in_{i,j}, & \text{for } R_{i,j}^{rep} \le r_{ij} < R_{i,j} \\ 0, & \text{for } R_{i,j} < r_{ij} \end{cases} \tag{8b}$$

where $\in_{ij}$ are the pair-wise interaction parameters,[6,26] and the interactions are counted for all pairs, except the first nearest neighbors along the chain. A strong soft-core repulsive energy of about 4kT can be used in the simulations. This term provides a lightly larger excluded volume for larger amino acids than that defined by the hard core. The values of the cut-off distances R_{ij}^{rep} and R_{ij} are given in Table I, below. The values of R_{ij} were adjusted to approximately mimic the contact distances employed in the derivation of binary interactions parameters.[20] Here, a "native" interaction scale as described by Skolnick, *et al.*[20]

TABLE I. Compilation of Pairwise Cut-off Distances in Angstroms

A_i	A_j	R_{ij}^{rep}	R_{ij} (attractive)[a]	R_{ij} (repulsive)
Small[b]	Small	4.35	7.03	6.32
Small	Large	4.57	7.03	6.32
Large	Large	4.83	7.50	7.03

[a] Attractive pair of amino acids.
[b] Small amino acids are: Gly, Ala, Ser, Cys, Val, Thr, Pro.

One-body burial interactions

To facilitate a rapid collapse of the model chain, a centro-symmetric, density regularizing term was used that is based on a statistical analysis of single domain proteins. This is the only term that uses the assumption that the target protein has a single domain. For some increase in computational cost, this term could be omitted. The radius of gyration of the protein is given by:

$$S = (N^{-1}\Sigma(r_{CM} - r_i)^2)^{1/2} \quad (9)$$

where r_{CM} is the position of the center of mass of the globule, and r_i is the position of the center of mass of the i-th side chain. The size of a single domain protein is strongly correlated with the number of residues, N, comprising the protein, in accordance with:

$$S = 1.52\ N^{0.38} \quad \text{in lattice units.} \quad (10)$$

The exponent 0.38, obtained from the statistical analysis of single domain globular proteins,[21] is very close to the value of 1/3 expected for a long, collapsed polymer chain.[22] The corresponding potential has the following form:[23]

$$E_b = \varepsilon_b\ \Sigma|m_{o.i} - m_i| \quad (11)$$

where $m_{o\,i}$ is the target number of amino acids in a given spherical shell centered at the protein's center of mass. There are three equal thickness shells within a distance S, and they contain somewhat more than half of the protein residues. The entire protein is essentially contained in a sphere of radius equal to 5/3 S. The value of the parameter ε_b was equal to 0.25-1.0 k_BT, depending on protein size. Larger proteins tend to exhibit a larger absolute deviation from the above target distribution of mass, and consequently, a lower penalty for such deviations should be employed.

To further enhance rapid collapse, those residues that are within a radius of 2/3 S (a very conservative estimate of the hydrophobic core of a single domain globular protein) contribute $\varepsilon_{KD}(i)/16$ to the total energy, where $\varepsilon_{KD}(i)$ is the Kyte-Doolittle hydrophobicity parameter of the i-th residue.[19,24] The scaling factor 1/16 is preferred. This potential (and its scaling with respect to other interactions) has very little effect on the folded structure, but it improves folding kinetics.

Multibody surface exposure term

Amino acid side groups have a different size and shape. Thus, when a given side chain is in contact with another amino acid, the fraction of its surface that is

covered depends on the identity of the contacting partner. Appropriate parameters reflecting this observation (*i.e.*, the surface coverage of particular types of side chains and associated statistical-type potential) could be derived from the statistics of known protein structures. In the present algorithm, each residue can have 30 surface contact points. A subset of these contact points becomes occupied upon contact with other side chains or main chain Cα atoms. The Cα atom positions are approximated from the positions of three consecutive side chain beads and have their own excluded volume and contribution to surface coverage. Due to "shadowing," *i.e.*, one residue being covered by another, some contact points could be multiply occupied by different residues (usually 1 or 2, or sometimes 3, but very rarely 4 or more). The fraction of occupied surface points defines the fraction of buried area of a given side chain. The total energy of a model protein is computed as:

$$E_{surface} = \varepsilon_S \Sigma\, E_b(A_i, a_i) \qquad (12)$$

where a_i is the covered fraction of sites of amino acid side chain A_i and E_b (A_i, a_i) is the statistical potential for amino acids A_i that are covered by a_i contact points, *i.e.*, its coverage fraction is a/30, when the number of contact points is 30. The reference state for this statistical potential is "an average" amino acid with average (over structural database) coverage. One scaling factor ε_S for this term has been determined to be 0.25, although other scaling can be used.

The above approach to the hydrophobic interactions allows suppression of previously employed centro-symmetric one-body potentials[6] and thereby opens up the present approach to multi-domain and multi-meric proteins. In this example, both models of mean field hydrophobic interactions were used in parallel.

The force field designed for this model is entirely of a "knowledge-based" origin. Some terms, such as the generic short- and long-range potentials, provide a bias toward protein-like short- and long-range correlations in the model chain. These potentials generalize regularities seen in native structures of all globular